mathematics

MDPI

Review

A Survey on Deep Transfer Learning and Beyond

Fuchao Yu, Xianchao Xiu and Yunhui Li *

School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China
* Correspondence: liyunhui@shu.edu.cn

Abstract: Deep transfer learning (DTL), which incorporates new ideas from deep neural networks into transfer learning (TL), has achieved excellent success in computer vision, text classification, behavior recognition, and natural language processing. As a branch of machine learning, DTL applies end-to-end learning to overcome the drawback of traditional machine learning that regards each dataset individually. Although some valuable and impressive general surveys exist on TL, special attention and recent advances in DTL are lacking. In this survey, we first review more than 50 representative approaches of DTL in the last decade and systematically summarize them into four categories. In particular, we further divide each category into subcategories according to models, functions, and operation objects. In addition, we discuss recent advances in TL in other fields and unsupervised TL. Finally, we provide some possible and exciting future research directions.

Keywords: deep transfer learning (DTL); domain adaptation; machine learning; neural networks

MSC: 68T05; 68T07; 68T09; 68T20

Citation: Yu, F.; Xiu, X.; Li, Y. A Survey on Deep Transfer Learning and Beyond. *Mathematics* **2022**, *10*, 3619. https://doi.org/10.3390/math10193619

Academic Editor: Ripon Kumar Chakrabortty, María Purificación Galindo Villardón, Jakub Nalepa

Received: 8 August 2022
Accepted: 29 September 2022
Published: 3 October 2022

1. Introduction

Machine learning has been successfully applied in many fields such as face recognition [1,2], autonomous driving [3], and smart healthcare [4]. Machine learning often emphasizes that training and testing data come from the same dataset and share consistent feature distributions. However, the consistency cannot be guaranteed in practical applications. In addition, as the acquired data become larger and more complicated, several problems arise, such as few annotations in datasets, poor computational capability of devices, and model generalization with limited data. For instance, millions of images in image processing may be contained in a dataset [5]. Labeling these images is an expensive and time-consuming task. A large amount of image data and a relatively small number of labels have triggered the contradiction between the large amount of data and the few labels and the contradiction between the large amount of data and the weak computing capability. Transfer learning (TL) has been proven to be efficient in solving the above problems. In addition, many researchers have demonstrated the theoretical viability of TL; see [6–13]. For example, Wang et al. [6] investigated model complexity and learning algorithm stability to derive TL theoretical bounds, Phung et al. [10] developed efficient algorithms of domain-invariant learning, and Wu et al. [12] described these from the perspective of information theory. A large number of TL-related approaches have been proposed, as can be seen in Figure 1.

TL reapplies the learned knowledge on source domains to achieve good performance on different but related target domains [14,15]. Next, we give some definitions of TL. A domain can be represented formally as $\mathcal{D} = \{\mathcal{X}, P(\boldsymbol{X})\}$, where $\mathcal{X}$ denotes a feature space and $P(\boldsymbol{X})$ denotes a marginal distribution for $\boldsymbol{X} = [\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^n] \in \mathbb{R}^{m \times n}$. For a specific domain $\mathcal{D} = \{\mathcal{X}, P(\boldsymbol{X})\}$, a task can be represented formally as $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$, where $\mathcal{Y}$ denotes a label space and $f(\cdot)$ denotes a decision function. Pan et al. [14] provided a definition of TL: given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, TL aims to help improve the learning of the decision function $f(\cdot)$ in $\mathcal{D}_T$

using knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$. As shown in Figure 2, we assume that Amazon and DSLR (digital single-lens reflex) are the source domain $\mathcal{D}_S$ and the target domain $\mathcal{D}_T$. We train a classification model to complete the category-level object detection $\mathcal{T}_S$ in the Amazon dataset. Transferring the trained parameters (knowledge) in the model to a new model will reduce the training cost and improve $\mathcal{T}_T$ in the DSLR dataset.

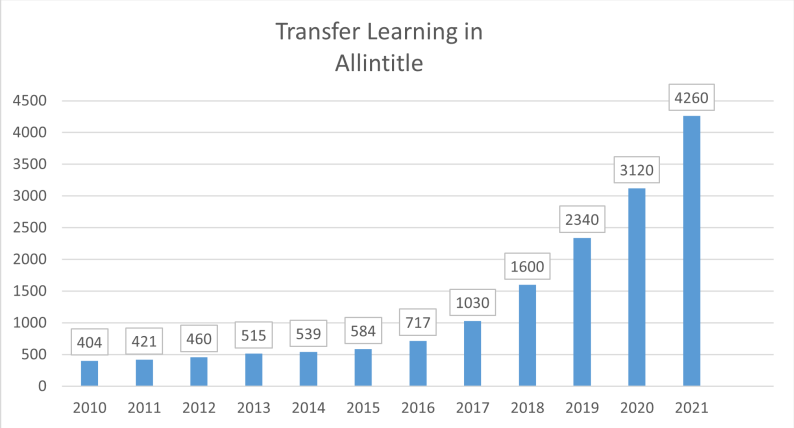

Figure 1. The development trend of TL from 2010 to 2021.

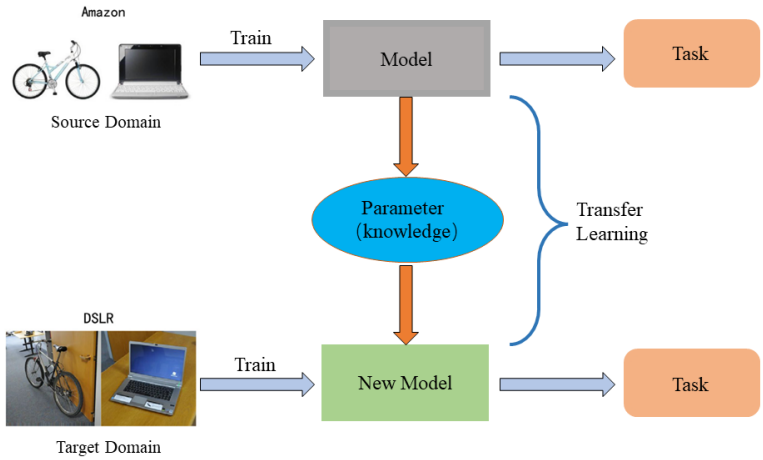

Figure 2. Intuitive explanation of TL.

According to [14–17], TL approaches can be categorized into four types: instance-based TL, model-based TL, feature-based TL, and relational-based TL. We provide a brief review of the four types as follows. (i) Instance-based TL completes the transfer by assigning different weights to different instances. A meaningful approach is using the ratio of source-domain and target-domain instances as sample weights [18–20]. Another method is the kernel mean matching approach [21], which matches the means between the source-domain and the target-domain instances in a reproducing kernel Hilbert space. (ii) Feature-based TL completes the transfer by transforming the features of different domains. One representative is the statistical feature transformation approach [22–24], which minimizes the distribution difference between source domains and target domains by statistical techniques. Another notable method is the geometric feature transformation approach [25–27], which implicitly aligns feature spaces between source domains and target domains by transforming features. (iii) Model-based TL completes the transfer by building models with shared parameters. These studies are broadly divided into two categories: the knowledge transfer based on shared model components and the regularization knowledge transfer based on a support vector machine (SVM). The former learns target-domain models by sharing source-domain models or hyperparameters [28,29]. The latter prevents overfitting by constraining hyperparameters with regularization terms [30,31]. (iv) Relational-based

TL completes the transfer by constructing a logical mapping relationship of source domains and target domains. It is assumed that the logical relationship between source and target domains has a common pattern. Thus, the logical relationship or rules learned in source domains can be transferred to target domains. One popular approach is the first-order Markov logic network [32].

With the development of deep neural networks (DNN) [33–36], many researchers have suggested integrating deep learning techniques with TL, thereby sharing both the advantages of deep learning and TL. Thus, a huge amount of deep transfer learning (DTL) frameworks have been constructed and have been shown to be promising. It should be noted that there exists a transition from TL to DTL, i.e., incomplete DTL. Specifically, this incompleteness lies in the fact that DNN simply act as feature extractors which are combined with shallow approaches [37–43]. For example, Donahue et al. [39] proposed the deep convolutional activation feature (DeCAF) for generic visual recognition. This work was to acquire generic features in the source domain by DNN in a fully supervised manner. After learning enough generalization-competent features, a simple linear classifier is used to handle this task in the target domain with no or few labels. Csurka et al. [37] used the maximum mean difference (MMD) to compensate for the differences between domains to improve the extraction performance of DNN. Then, they adopted shallow approaches to complete the classification. Li et al. [38] studied a low-rank parameterized convolutional neural network that extracts common features in source and target domains to accomplish TL. Although these works have achieved relatively good performance due to the fact that DNN can learn excellent transferable representations, the two-step learning process causes the accumulation of errors to affect the final accuracy. As suggested in [44], end-to-end learning approaches can overcome this shortcoming. Therefore, increasing numbers of researchers have explored DNN architectures to construct DTL models. Figure 3 provides an overview of these approaches in chronological order.

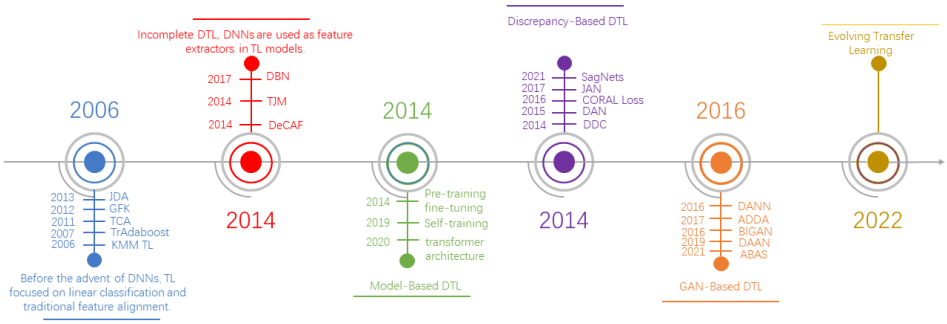

Figure 3. Timeline diagram of TL development.

This survey provides a comprehensive review of the recent development of DTL. The main contributions of this survey are summarized as follows.

- We introduce over 50 representative approaches of DTL and systematically summarize them into four categories and further subcategories; see Figure 4.
- We present frontier advances in the application of DTL and recent advances in unsupervised TL.
- We provide some potential research directions that can give a good reference for promoting future work in this field.

This paper is structured as follows. Sections 2–5 summarize each of the four approaches: DTL, including model-based DTL, discrepancy-based DTL, GAN-based DTL, and relational-based DTL; see Table 1. Section 6 gives extensions and additions of DTL. Section 7 concludes this paper with suggestions for future research. In addition, open-source codes and datasets of DTL approaches are presented in Appendix A.

Notation: In this paper, all spaces are indicated by calligraphic uppercase letters, i.e., $\mathcal{X}$; all matrices are denoted by bold uppercase italic letters, i.e., $\boldsymbol{X}$; all vectors are represented by

bold lowercase italic letters, i.e., $\boldsymbol{x}$; and all scalars are defined as lowercase italic letters, i.e., λ. Moreover, all variable indices, such as i, j, are denoted by italic superscripts, while specific names, such as source domain S, target domain T, label predictor y, domain classifier d, feature extractor f, are upright and denoted in subscripts. Let $\mathbb{R}^{m\times n}$ characterize the set of all $m \times n$ matrices. Denote $\boldsymbol{X} = [\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^n] \in \mathbb{R}^{m\times n}$ be the input samples matrix, where m is the number of feature dimension and n is the number of samples.

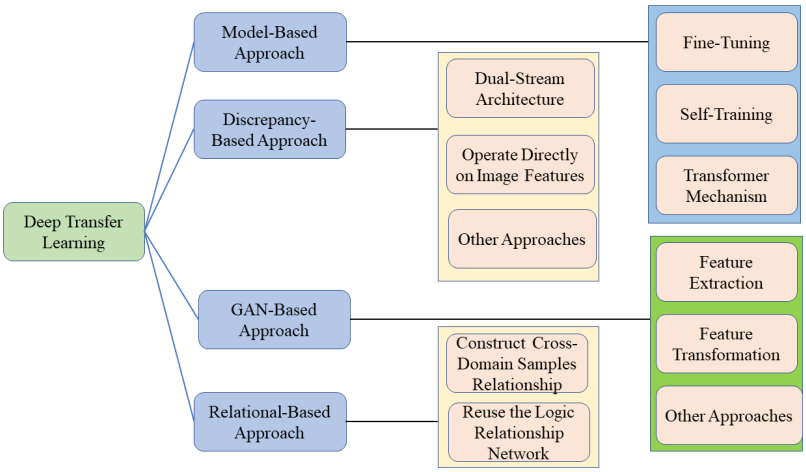

Figure 4. Categorizations of DTL.

Table 1. A brief summary of DTL approaches.

DTL Approaches	Subcategories	Brief Description
Model-Based DTL	Fine-Tuning [43,45–50] Self-Training [51–53] Transformer Mechanism [54–58]	Share and fine-tune the parameters of deep learning models
Discrepancy-Based DTL	Dual-Stream Architecture [59–72] Operate on Image Features [73–75]	Reduce feature discrepancies between source and target domains by DNN
GAN-Based DTL	Feature Extraction [76–81] Feature Transformation [82–90]	Extract domain invariant features by generative adversarial networks
Relational-Based DTL	Cross-Domain Relationship [91,92] Logical Networks [93,94]	Construct relationship using cross-domain relationship or logical networks

2. Model-Based DTL

In this section, model-based DTL approaches are summarized into three categories. One is the fine-tuning model. which fine-tunes the parameters of source-domain networks to achieve good performance in target domains [43,45–50]. The second is the self-training approach, which is adopted to overcome the limitations of the fine-tuning model in the case of data enhancement and annotation increases [51–53]. The third is the transformer-based architecture, which introduces the attention mechanism in the image recognition field [54–58]. We give a brief summary of model-based DTL approaches in Table 2.

Table 2. A brief summary of model-based DTL approaches.

Model-Based DTL Approaches	Representatives	Brief Description
Fine-Tuning	Yosinski et al. [45], DLID [43], Rozantsev et al. [47]	Reuse the different layer parameters of DNN trained in source domains
Self-Training	He et al. [51], Xie et al. [52], Zoph et al. [53]	Enhance model performance by using predicted pseudo-labels and noise
Transformer Mechanism	BEiT [55], TVT [57], Xu et al. [58]	Share and fine-tune parameters of transformer in target domains

2.1. Fine-Tuning

Fine-tuning is one of the earliest attempts at model-based DTL, which can be traced back to the interpolation path approach proposed by DLID: deep learning for domain adaptation by interpolating between domains [43]. Yosinski et al. [45] investigated the transferability of AlexNet in 2014. The researchers divided 1000 classification datasets into two equal parts: A and B. They trained two networks for the two datasets and fine-tuned the first seven layers of the network one by one to investigate the role of different layers in the model transfer process. The experimental results show that "transfer plus fine-tuning" leads to the best performance. The following conclusions are drawn from this experiment: the first three layers of AlexNet are general features that facilitate a transfer, and adding fine-tuning to the neural network can overcome the variability of the data to improve the network performance. Chu et al. [46] reached the same conclusion based on this experiment by considering the effect of dataset bias and the number of target-domain data markers in this experiment. These studies illustrate that the choice of the fine-tuning layer affects the model performance, which provides the basis for subsequent improvements in fine-tuning approaches.

Although the approach of fine-tuning parameters of the first few layers has subsequently been widely used, there was no clear basis for determining sharing layers. Until 2018, Rozantsev et al. [47] proposed a deep domain adaptation approach, which selectively shared and restricted parameters of different layers. The approach introduces a maximum mean difference (MMD) loss function in a dual-stream structure to measure the same layers of neural networks trained simultaneously in the target and source domains. Then, the weights for the restricted layers with large MMD losses are regularized so that the parameters satisfy a certain linear relationship. The objective function optimized in this approach by minimizing the loss function is [47]

$$L(\boldsymbol{\Theta}_{\mathrm{S}}, \boldsymbol{\Theta}_{\mathrm{T}} | \boldsymbol{X}_{\mathrm{S}}, Y_{\mathrm{S}}, \boldsymbol{X}_{\mathrm{T}}, Y_{\mathrm{T}}) = L_{\mathrm{S}} + L_{\mathrm{T}} + L_{\mathrm{W}} + L_{\mathrm{DD}}, \tag{1}$$

where loss functions are

$$\begin{aligned} L_{\mathrm{S}} &= \frac{1}{n_{\mathrm{S}}} \sum_{i=1}^{n_{\mathrm{S}}} c\left(\boldsymbol{\Theta}_{\mathrm{S}} \middle| \boldsymbol{x}_{\mathrm{S}}^{i}, y_{\mathrm{S}}^{i}\right), \\ L_{\mathrm{T}} &= \frac{1}{n_{\mathrm{T}}} \sum_{j=1}^{n_{\mathrm{T}}} c\left(\boldsymbol{\Theta}_{\mathrm{T}} \middle| \boldsymbol{x}_{\mathrm{T}}^{j}, y_{\mathrm{T}}^{j}\right), \\ L_{\mathrm{w}} &= \lambda_{\mathrm{w}} \sum_{k \in \Omega} r_{\mathrm{w}}\left(\theta_{\mathrm{S}}^{k}, \theta_{\mathrm{T}}^{k}\right), \\ L_{\mathrm{DD}} &= \lambda_{\mathrm{u}} r_{\mathrm{u}}(\boldsymbol{\Theta}_{\mathrm{S}}, \boldsymbol{\Theta}_{\mathrm{T}} | \boldsymbol{X}_{\mathrm{S}}, \boldsymbol{X}_{\mathrm{T}}). \end{aligned} \tag{2}$$

where $\boldsymbol{\Theta}_{\mathrm{S}} = \{\theta_{\mathrm{S}}^{i}\}$ and $\boldsymbol{\Theta}_{\mathrm{T}} = \{\theta_{\mathrm{T}}^{j}\}$ denote the parameters of all layers in the source and target domains. $\boldsymbol{X}_{\mathrm{S}} = \{\boldsymbol{x}_{\mathrm{S}}^{i}\}_{i=1}^{n_{\mathrm{S}}}$ and $\boldsymbol{X}_{\mathrm{T}} = \{\boldsymbol{x}_{\mathrm{T}}^{j}\}_{j=1}^{n_{\mathrm{T}}}$ are the sets of samples from the source and target domain, respectively. $y_{\mathrm{S}}^{i} \in Y_{\mathrm{S}}$ and $y_{\mathrm{T}}^{j} \in Y_{\mathrm{T}}$ are the label set corresponding to $\boldsymbol{x}_{\mathrm{S}}^{i}$ and $\boldsymbol{x}_{\mathrm{T}}^{j}$. n_{S} and n_{T} denote the number of samples in the source and target domains, respectively.

$c(\cdot)$ is a standard classification loss. Furthermore, $r_{\mathrm{w}}(\cdot)$ and $r_{\mathrm{u}}(\cdot)$ are the weight and unsupervised regularizers. The regularizer L_{W} acts on the set Ω of indices of the layers whose parameters are not shared and represents the loss of the corresponding layer of the two streams. L_{DD} encodes the domain differences to produce a similar distribution between the source and target domains' data representations. It is assumed that the target samples are ordered such that only the first n_{T} has valid labels where $n_{\mathrm{T}} = 0$ in an unsupervised scenario. Since no target-domain labels are available, the optimization function $L_{\mathrm{T}} = 0$ for the standard classification loss in the target domain. These regularizers are weighted by coefficients λ_{w} and λ_{u}, respectively. More illustrations can be found in Figure 5.

After that, Rozantsev et al. [48] and Guo et al. [49] proposed auxiliary residual networks with adaptive fine-tuning techniques to selectively freeze and adjust the parameters, respectively. The former used two-stream network structures similar to [47] with residual transformations and performed on each neural network layer to fit the target-domain data. The latter used different parameter-tuning strategies for different instances of the target-domain data. Note that the decision to freeze or fine-tune the parameters of the pre-trained network was generated based on the Gumbel SoftMax distribution. The latest fine-tuning research addresses the overfitting problem when the number of target datasets is small. Li et al. [50] proposed to interpolate between regularization and self-labeling, including layer-wise regularization, self label-correction, and label re-weighting.

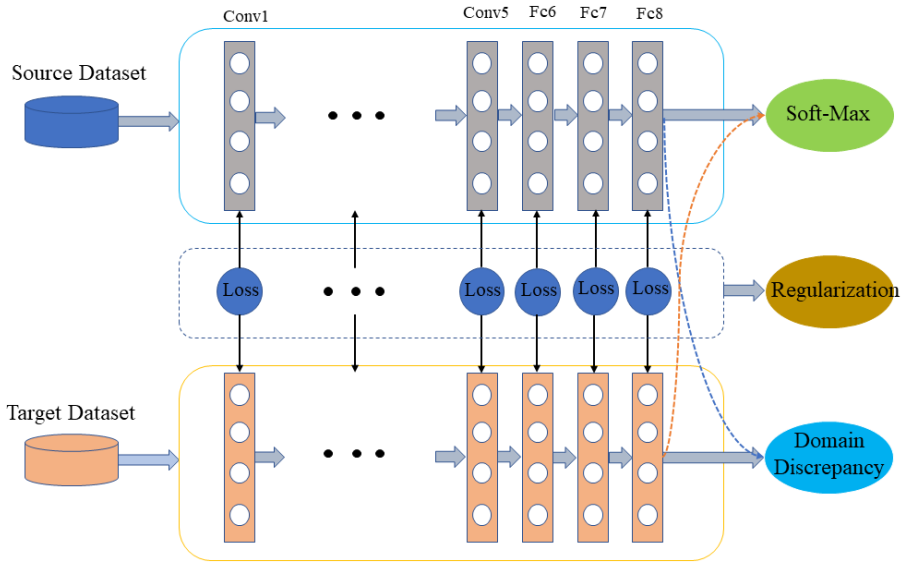

Figure 5. Two-stream optimized architecture of fine-tuning [47]. Here, Conv1–Conv5 denote convolutional layers and Fc6–Fc8 denote fully connected layers.

2.2. Self-Training

Although fine-tuning has achieved great success in some applications, the flaws of fine-tuning have also been identified by researchers. He et al. [51] first discovered limitations of the fine-tuning model when performing cross-dataset implementations of target detection and semantic segmentation tasks. It is found that the pre-trained model on the ImageNet dataset performed worse for the COCO dataset than the random initialization parameters approach, and ImageNet pre-training can accelerate convergence in the early stages but cannot provide regularization or improve accuracy in the final task. To overcome the disadvantage of fine-tuning, Xie et al. [52] proposed a self-training model for the weakly supervised domain adaptation problem, which uses a small number of labels from the source domain to TL across datasets. The training process is given as follows, with labeled images $\{(x^1, y^1), (x^2, y^2), \ldots, (x^n, y^n)\}$ and unlabeled images $\{\tilde{x}^1, \tilde{x}^2, \ldots, \tilde{x}^m\}$. θ is parameters in networks and denotes the networks.

- Train a teacher network θ which minimizes the cross entropy loss $L(\cdot)$ on partially labeled images, defined as

$$\frac{1}{n}\sum_{i=1}^{n} L\left(y^i, f\left(x^i, \theta\right)\right), \tag{3}$$

where $f(\cdot)$ is a prediction function which predicts labels using network parameter θ.
- The teacher network θ is used to predict the unlabeled images, and the predictions $\tilde{y}_i$ are used as pseudo-labels. The mathematical model is given by

$$\tilde{y}^i = f\left(\tilde{x}^i, \theta\right), \forall i = 1, 2, \ldots, m. \tag{4}$$

- Train a student network η which minimizes the cross entropy loss $L(\cdot)$ on labeled and pseudo-labeled samples as follows:

$$\frac{1}{n}\sum_{i=1}^{n} L\left(y^i, f_{\text{noised}}\left(x^i, \eta\right)\right) + \frac{1}{m}\sum_{i=1}^{m} L\left(\tilde{y}^i, f_{\text{noised}}\left(\tilde{x}^i, \eta\right)\right), \tag{5}$$

where $f_{\text{noised}}(\cdot)$ is a prediction function. Noises such as dropout, random depth, and data augmentation are added to enhance the representational power of the student network.
- The student network η is used as a new teacher network θ_*; then, return to step 2.

The self-training approach is trained on unlabeled datasets to obtain generalized data representations. The key to self-training is adding noise in the training process of the student network, which enhances the smoothness of the decision function in both labeled and unlabeled data to obtain higher performance than the teacher network. Network performance is continuously enhanced during multiple iterations.

Recently, Zoph et al. [53] compared the self-training with fine-tuning and set up three control experiments through data augmentation and data-annotation addition. It can be concluded from Figure 6 that (i) stronger data augmentation and more labeled data further reduce the value of fine-tuning. (ii) Unlike fine-tuning, self-training always contributes to the training accuracy at any data augmentation strength. (iii) Even when fine-tuning works, a strategy incorporating self-training can improve performance. Therefore, combining self-training with fine-tuning approaches can greatly improve the cross-domain learning performance, which has good prospects for future development.

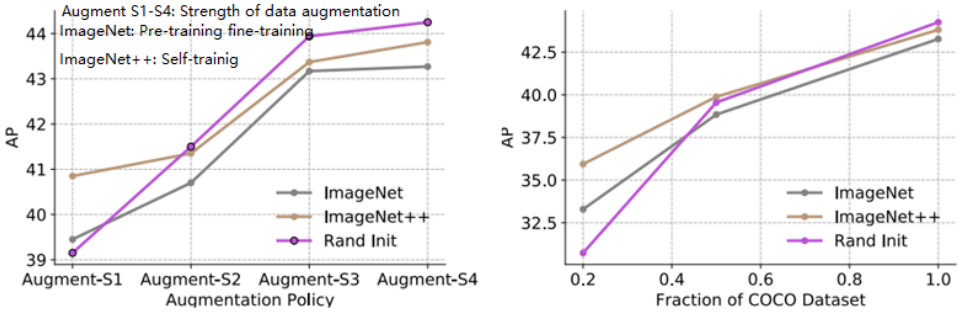

Figure 6. Experimental results of data augmentation (**left**) and data-label addition (**right**) [53].

2.3. Transformer Attention Mechanism

The transformer attention mechanism has been demonstrated to be promising in natural language processing (NLP) [95–98]. Dosovitskiy et al. [56] successfully introduced a transformer into computer vision by chunking and spreading images into one-dimensional vectors. This mechanism provides a new framework for fine-tuning, which is different from convolutional neural networks (CNN). Chen et al. [54] first implemented a fine-tuning architecture with a transformer. They proposed the image processing converter (IPT), which applies transformers to underlying computer vision tasks. For different

settings, transformer modules are shared, and only new head and tail structures need to be replaced according to the task requirements. Figure 7 shows the structure of the fine-tuning transformer. The IPT achieved state-of-the-art performance on several underlying visual tasks such as super-resolution, denoising, and rain sound removal. Bao et al. [55] proposed a self-supervised visual representation model, called Bert pre-training of image transformers (BEiT), which borrows from Bert in the field of NLP. At first, the data are pre-processed to two views: the original image and the image after randomly masking part of the image block. The goal of pre-training is to recover the original image based on the corrupted image by using a transformer. After pre-training BEiT, the model parameters on downstream tasks are fine-tuned directly by appending task layers to the pre-trained encoder. This approach allows the transformer to automatically acquire semantic region knowledge without markers so that the fine-tuning performance is greatly improved.

There are some researchers that are directly optimizing the model architecture. For example, Yang et al. [57] proposed the transferable visual transducer (TVT), which exploits the attention mechanism of the vision transformer (ViT) and the advantages of sequential images for knowledge transfer. They completed the TL by injecting the learned transferability into the attention block through a designed transferability adaptation module (TAM). Xu et al. [58] used cross-attention in transformers for feature alignment. They proposed a weight-sharing three-branch converter framework to apply self-attention and cross-attention for source-target feature learning and source-target domain alignment, respectively. However, the robustness and interpretability of model-based DTL need to be further investigated.

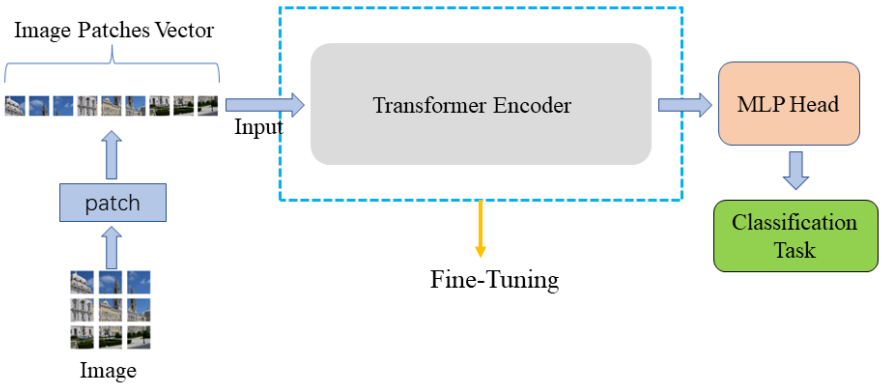

Figure 7. Fine -tuning transformer of IPT [57].

3. Discrepancy-Based DTL

The fine-tuning model often requires the network to contain a large amount of labeled sample data and a similar distribution of data features in the source and target domains, which cannot solve unsupervised and cross-domain problems well. Discrepancy-based DTL further explores the architecture of DNN and features of the source and target domains, which provides good solutions to the drawbacks of fine-tuning approaches. Dual-stream architectures [59–72] and approaches that operate directly on image features [73–75] are representative ways to perform discrepancy-based DTL. In addition, optimizing the network architecture [99] and improving the feature alignment [100–103] have also received attention from researchers in recent years. The essence of the above approaches is to minimize the feature differences between the source and target domain datasets. We give a brief summary of discrepancy-based DTL approaches in Table 3.

Table 3. A brief summary of discrepancy-based DTL approaches.

Discrepancy-Based DTL Approaches	Representatives	Brief Description
Dual-Stream Architecture	DDC [60], DAN [62], JAN [64]	Reduce domain differences by introducing adaptation layers in DNN
Operate Directly on Image Features	SagNets [73], Yoon et al. [74], Yu et al. [75]	Operate directly on features extracted by DNN to align differences
Other Approaches	Das et al. [99], Li et al. [101], Zhu et al. [103]	Optimize the network architecture and improve the feature alignment

3.1. Dual-Stream Architecture

The dual-stream structure [104] is currently an essential framework for discrepancy-based DTL, whose core is the reduction of domain differences by introducing adaptation layers in DNN. This end-to-end learning reduces the error accumulation of multi-step learning in traditional TL. The deep architecture takes original data as input, in which neural networks process data, extract features, and align the domain. Currently, neural networks such as AlexNet [33], VGG [34], GoogleNet [35], and ResNet [36] are used as stream models for dual-stream architectures in the field of image classification and detection. The idea of dual-stream architectures is to use the same neural network to simultaneously train the source and target domains. An adaptation layer is added into the network to minimize the differences between the two domains. In addition to the different networks, the main difference lies in the design of the adaptation layer and the use of loss functions. Maximum mean difference (MMD) loss is the most common alignment in dual-stream architecture, which is formulated as [105]

$$MMD(\boldsymbol{X}_{\mathrm{S}}, \boldsymbol{X}_{\mathrm{T}}) = \left\| \frac{1}{n_{\mathrm{S}}} \sum_{i=1}^{n_{\mathrm{S}}} \Phi\left(\boldsymbol{x}_{\mathrm{S}}^{i}\right) - \frac{1}{n_{\mathrm{T}}} \sum_{j=1}^{n_{\mathrm{T}}} \Phi\left(\boldsymbol{x}_{\mathrm{T}}^{j}\right) \right\|_{\mathcal{H}}^{2}, \tag{6}$$

where $\boldsymbol{x}_{\mathrm{S}}^{i} \in \boldsymbol{X}_{\mathrm{S}}$ and $\boldsymbol{x}_{\mathrm{T}}^{j} \in \boldsymbol{X}_{\mathrm{T}}$ are the source and target domain samples, respectively. n_{S} and n_{T} denote the number of samples in the source and target domains, respectively. $\Phi(\cdot)$ defines a mapping from raw data to the reproducing kernel Hilbert space (RKHS). $\mathcal{H}$ indicates that the distance is metricized in RKHS.

Ghifary et al. [59] introduced MMD loss in neural networks to improve the domain adaptation performance of the network, which is the first use of MMD in DNN. In the research of the adaptation layer, MMD loss is broadly used to align different domains. Tzeng et al. [60] started the study of dual-stream architecture in TL and proposed the deep domain confusion (DDC) approach. The design of this architecture is shown in the gray part of Figure 8. They optimized a CNN architecture for classification loss $L_{\mathrm{C}}(\cdot)$ and domain loss $L_{\mathrm{MMD}}(\cdot)$. An adaptation layer is introduced in the previous layer of the classifier, and a domain confusion loss is computed from the output of the adaptation layer. The MMD distance between the source and target domains features is used as the domain loss, which is minimized to reduce the difference between the source and target domains. The loss function $L_{\mathrm{DDC}}(\cdot)$ of this approach can be expressed as

$$L_{\mathrm{DDC}} = L_{\mathrm{C}} + \lambda L_{\mathrm{MMD}}, \tag{7}$$

where $L_{\mathrm{C}}(\cdot)$ and $L_{\mathrm{MMD}}(\cdot)$ denote the classification loss in the source domain and domain loss, respectively. The hyperparameter λ is used to determine the influence of domain confusion on the optimization. They jointly optimized this loss function by minimizing the classification loss and the MMD loss to maximize the domain confusion loss.

After that, Tzeng et al. [61] improved the DDC approach and designed a new CNN structure. The difference with DDC is that the last layer of the target domain network can output soft label loss. The three losses are optimized simultaneously to achieve cross-domain and cross-task recognition. Thus, the loss function can be further improved by

$$L = L_{\mathrm{DDC}} + vL_{\mathrm{soft}}, \tag{8}$$

where $L_{\mathrm{soft}}(\cdot)$ is the soft label loss in the target domain and v determines the soft label weights. This loss trains the network parameters to produce a "soft label" activation that matches the average output distribution of source examples on a network trained to classify source data. For details, please refer to Figure 8. Furthermore, Long et al. [62] proposed a deep adaptation network (DAN) architecture by adding three adaptive layers simultaneously to the first three layers of the classifier for feature constraint, which can match both low-order moments and high-order moments.

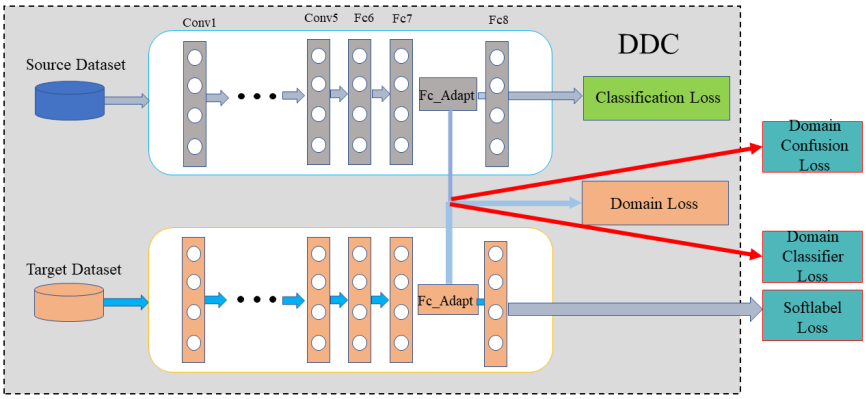

Figure 8. Schematic diagram of the improved DDC (gray) architecture [60,61]. Here, Fc_Adapt is an adaptation layer.

The adaptation layer introduced above can be considered a marginal domain adaptation approach. The following presents the conditional distribution adaptation, joint distribution adaptation, and dynamic distribution adaptation approaches. Zhu et al. [63] proposed the depth subdomain adaptation network (DSAN), which is a conditional distribution adaptation approach. DSAN focuses on subdomain adaptation and learns a transfer network by aligning the relevant subdomain distribution of domain-specific layer activations across domains by the local maximum mean difference (LMMD). Long et al. [64] proposed the joint adaptation network (JAN), which is a representative approach to joint distribution adaptation; see Figure 9. JAN learns a transfer network by aligning the joint distribution of multiple domain-specific layers across domains based on the joint maximum mean difference (JMMD). They adopted an adversarial training strategy to maximize JMMD, which leads to more distinguishable distributions in the source and target domains. Wang et al. [65] proposed a dynamic distribution adaptation method (DDAN) by improving the above approach. The same network structure as the JAN and DAN is used. In this work, dynamic adaptation units are embedded in the feature layer to introduce dynamic factors that dynamically adjust the weight of the marginal and conditional distributions.

The dual-stream architecture has been widely used in the field of multi-representation learning. Zhu et al. [67] proposed the multi-representation adaptive network (MRAN) approach. MRAN obtains multiple representations of the original image, and then the feature alignment is performed in different feature spaces separately to improve the accuracy of cross-domain image recognition tasks.

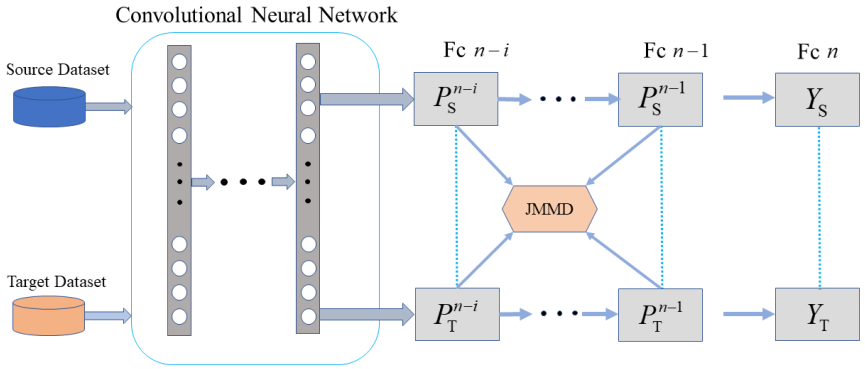

Figure 9. The architecture of JAN [64]. Here, Fc n is an output layer (the last layer) of the neural network. The joint distributions of the deep network activations $P_S^{n-i}, \ldots, P_S^{n-1}$ and $P_T^{n-i}, \ldots, P_T^{n-1}$ in these layers are adapted by JMMD minimization.

The DTL with dual-stream architecture presented above uses MMD loss as a rule, and the following are other forms of rules in dual-stream architecture. Correlation Alignment (CORAL) loss [68] is a well-known approach for feature alignment in dual-stream architecture, which proposes a simple and effective unsupervised adaptation that extends the linear transformation of the traditional CORAL to a nonlinear transformation. The approach accomplishes the image classification problem when the target domain is unlabeled by optimizing the source domain classification loss and CORAL loss to align the second-order statistics of domain distributions. The central moment discrepancy (CMD) approach [66] is different from the standard matching distribution approach of MMD. CMD matches the higher-order central moments of the probability distribution by sequential moment differences, which provides a new distance function for domain-invariant representation learning. The adaptive batch normalization (AdaBN) approach [69] modulates the statistical information from the source domain to the target domain in all batch normalization layers of DNN and normalizes the data to achieve the DTL. In addition, Wasserstein distance [70] is used to minimize the feature distribution differences and the transportation cost in the optimal transportation-based DTL [71,72].

3.2. Operation with Image Features

The dual-stream architecture requires various loss functions to align the discrepancies that occur when two domains are trained for the network. However, various loss functions suffer from low robustness and do not adapt well to differences in the distribution of the source and target domains. The approaches described below operate directly on image features to compensate for differences between features.

A recent study found that the main reason for the inability of deep learning to transfer across datasets in image classification tasks effectively is that CNN is more sensitive to image texture features. Therefore, the significant differences of image texture features in different datasets are one of the main reasons that prevent TL from performing effectively. Nam et al. [73] proposed style-agnostic networks (SagNets) to achieve separation of style encoding from image content for reducing domain bias. The feature extractor of SagNets (see Figure 10) extracts not only the content of the image but also the image style. In the content-biased network, the styles are randomly initialized by adaptive instance normalization (AdaIN) to make this network focus on the image content. In the style-biased network, the opposite is true. Yoon et al. [74] proposed the knowledge distillation approach, which is the latest research on style features. They generated an assistant feature by transferring an intermediate style between labeled and unlabeled samples. They then trained a TL model by minimizing the output discrepancy between the unlabeled samples and the assistant. In addition, Yu et al. [75] combined meta-learning for learning distribution matching in a data-driven manner to reduce inductive bias and proposed an approach called learning to

match (L2M). L2M is a versatile framework that has shown excellent performance in the application of transfer of pneumonia to COVID-19 chest X-ray images.

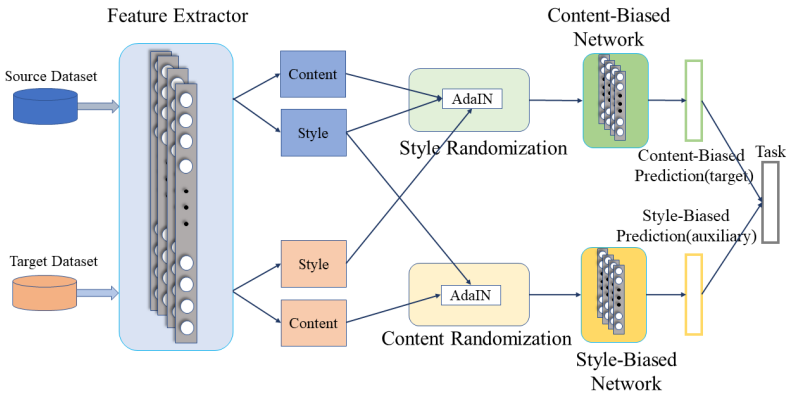

Figure 10. Framework of the SagNets [74].

3.3. Other Approaches of Feature Transfer

Apart from the above two types, some other approaches to feature transfer are described below. Das et al. [99] adapted existing domain adaptation methods to two new methods for the single rare class setting: DeerDANN, based on the Domain-Adversarial Neural Network (DANN), and DeerCORAL, based on deep correlation alignment (Deep CORAL) architectures. The two architectures augment the under-represented classes with synthetic samples, alleviating the lower classification performance for rare classes in both datasets. Li et al. [101] realized TL by using out-of-distribution detection (OOD) approaches in DNN. They trained the model to transfer domain perturbations and achieved better robustness against potential domain shifts by modeling the uncertainty of domain shifts with synthesized feature statistics during training. Although aligning local domains as closely as possible can make the connection between each neighboring domain stronger, it is worse for the alignment of distant domains. Xu et al. [100] proposed to use topology to accomplish domain adaptation. This approach reduces the effect of uniform alignment by using domain maps to encode neighboring domains. For multi-source domains learning, Ghifary et al. [102] proposed an encoder corresponding to multiple decoders. The main idea is to extract features shared across domains by a training autoencoder that reconstructs the data from different domains. The input is image data, and the output is a reconstruction of all domain analogs to that image. Zhu et al. [103] proposed a new framework with two alignment phases, which extracts domain-invariant representations of all domains by aligning the distributions of the source and target domain pairs in the common feature space.

4. GAN-Based DTL

With the great success of the generative adversarial network (GAN) [99] in image processing, researchers have attempted to incorporate the idea of GAN with TL to improve cross-domain learning. GAN is composed of two sub-networks (multilayer perception), including a generator $G(\cdot)$ and a discriminator $D(\cdot)$. The objective function is given by

$$\min_{G}\max_{D} E_{\boldsymbol{x}\sim P_{\mathrm{x}}}(\log(D(\boldsymbol{x})) + E_{\boldsymbol{x}\sim P_{\mathrm{z}}}\log(1 - D(G(\boldsymbol{z})))), \tag{9}$$

where $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$ denote samples from P_{x} and P_{z}, respectively, $G(\cdot)$ learns the mapping from a priori distributions P_{z} to true data distributions P_{true}, and $D(\cdot)$ denotes probability that the input comes from the true data. $G(\cdot)$ and $D(\cdot)$ compete with each other to complete adversarial training. $E(\cdot)$ is the expected probability of different distributions. When the game reaches equilibrium, the generator can generate true-looking samples.

Adversarial transfer learning (ATL) can "translate" between source-domain samples and target-domain samples while preserving the original label information for TL. ATL reduces domain discrepancies by solving the max-min game problem, which differs from the above approaches. Below, we review feature extraction approaches, feature transformation approaches, and other impressive approaches; see Table 4.

Table 4. A brief summary of GAN-based DTL approaches.

GAN-Based DTL Approaches	Representatives	Brief Description
Feature Extraction Approach	DANN [76], DAAN [79], Long et al. [77]	Extract invariant features from the source and target domains in adversarial training
Feature Transformation Approach	ADDA [84], SimGAN [85], Zhu et al. [81]	Transform the features for reducing the domain bias by adversarial training
Other Approaches	ALI [106], Ma et al. [107], Kang et al. [108]	—

4.1. Feature Extraction Approach

The feature extraction approach of ATL extracts invariant features from the source domain and the target domain in adversarial training for TL. Ganin et al. [76] proposed the domain-adversarial neural network (DANN), which is the first approach to add an adversarial mechanism to the training of neural networks. The approach learns domain-invariant features using a feature extractor and domain discriminator competing with each other. At this time, the features extracted from the source and target domains become increasingly similar so that classification tasks can be completed in the target domain by a classifier from the source domain. Figure 11 shows the architecture of DANN. The network requires the presence of labels on the source domain data to obtain the classification loss and the source and target domain data to be separable for obtaining the domain classification loss. It passes the two losses through the gradient reversal layer (GRL) to the feature extractor for back-propagation optimization. The final goal is to make it impossible for $G_{\mathrm{d}}(\theta_{\mathrm{d}})$ to discriminate between the features passed by the feature extractor. The parameters θ_{f} and θ_{y} are optimized in this process by minimizing the classification loss for the feature extractor

$$\left(\hat{\theta}_{\mathrm{f}}, \hat{\theta}_{\mathrm{y}}\right) = \arg\min_{\theta_{\mathrm{f}}, \theta_{\mathrm{y}}} E\left(\theta_{\mathrm{f}}, \theta_{\mathrm{y}}, \theta_{\mathrm{d}}\right). \tag{10}$$

Then, the loss of $G_{\mathrm{d}}(\cdot)$ is maximized to optimize the parameters θ_{d}.

$$\left(\hat{\theta}_{\mathrm{d}}\right) = \arg\max_{\theta_{\mathrm{d}}} E\left(\theta_{\mathrm{f}}, \theta_{\mathrm{y}}, \theta_{\mathrm{d}}\right). \tag{11}$$

The DANN objective function can be obtained by adding a gradient inversion layer and merging the two training processes:

$$E\left(\theta_{\mathrm{f}}, \theta_{\mathrm{y}}, \hat{\theta}_{\mathrm{d}}\right) = \sum_{\boldsymbol{x} \in D_{\mathrm{S}}} L_{\mathrm{y}}\left(G_{\mathrm{y}}(G_{\mathrm{f}}(\boldsymbol{x})), y\right) - \lambda \sum_{\boldsymbol{x} \in D_{\mathrm{S}} \cup D_{\mathrm{T}}} L_{\mathrm{d}}(G_{\mathrm{d}}(G_{\mathrm{f}}(\boldsymbol{x})), d), \tag{12}$$

where $L_{\mathrm{y}}(\cdot)$ and $L_{\mathrm{d}}(\cdot)$ denote classification loss and discriminator loss, respectively. Note that d is the label of domains: when the data come from the source domain, $d = 0$; otherwise $d = 1$. $\boldsymbol{x} \in \mathbb{R}^d$ and y are the input images and corresponding labels.

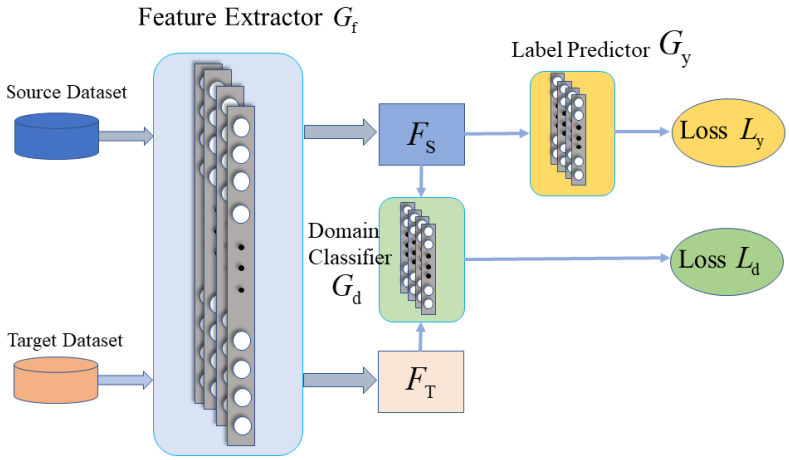

Figure 11. Architecture of DANN [76]. Here, DANN consists of three sub-networks: a feature extractor shared between domains $G_f(\theta_f)$, a label predictor for source domain category classification $G_y(\theta_y)$, and a domain discriminator to determine the origin of features $G_d(\theta_d)$. F_S and F_T are features from the source and target domains, which can be understood as mappings of the original input. L_y and L_d denote the classification loss function in the source domain and the domain classification loss function. Note that, f, d and y are labels of feature extractor, domain classifier, and label predictor, respectively.

The discriminator in DANN receives overall features of the source and target domains, equivalent to directly optimizing the difference between the distribution of $P_S(\boldsymbol{X})$ and $P_T(\boldsymbol{X})$, which considers the overall distribution of data features and ignores the correlation between categories. Long et al. [77] proposed conditional adversarial domain adaptation for the improvement of the DANN, which adapts both features and categories to obtain the relationship between deep features. The approach uses multilinear mapping to optimize the GAN, which somewhat improves the negative migration. Pei et al. [78] proposed the multiple adversarial domain adaptation (MADA) based on DANN, which captures multi-modal structures to achieve fine-grained alignment of different data distributions based on multiple domain discriminators.

Inspired by the above work, Yu et al. [79] further optimized the DANN and proposed the dynamic adversarial adaptation network (DAAN) for TL. They introduced adaptive factors in the design of domain loss to dynamically and quantitatively evaluate the contribution of both marginal distribution and conditional distribution decisions to adversarial learning. Figure 12 shows the overview of DAAN. Feature extraction is performed by the depth feature extractor (blue). The features are input in the calculation of domain loss, and the dynamic measurement factor ω is updated by dynamically measuring the weights of the overall feature domain classifier (purple) and multiple local feature classifiers (green). The classifier (orange) uses a stable DNN to solve the classification in the source domain. In addition, data augmentation is also a way to learn domain invariant features. Xu et al. [80] proposed incorporating domain confusion into TL to learn the common features of source and target domain data.

4.2. Feature Transformation Approach

Feature transformation is an important approach of ATL, which transforms or aligns the features as well as reducing the domain bias by adversarial training. Domain mapping is a representative approach for feature transformation in ATL. Adversarial discriminative domain adaptation (ADDA) [84] is an approach that combines discriminative modeling, non-shared weights, and GAN losses. ADDA first learns discriminator representation by using labels in the source domain to produce domain adversarial loss, which is used to learn asymmetric mapping for mapping target-domain data to separate encoding in the same space as the source domain. Distinct from mapping target-domain data in ADDA,

simulated GAN (SimGAN) [85] translates the source domain samples to the target domain for learning recognition classifiers available on two domains. Similar to SimGAN, Volpi et al. [83] used adversarial learning to translate labeled source domain data into target-domain samples while retaining the source-domain labels in the process. Domain mapping can also be used to learn domain invariant features. Zhu et al. [81] proposed cycle-consistent adversarial networks, which are trained by measuring the differences between the data after the source–target–source mapping and the original data. Feature alignment by adversarial training has also attracted the attention of researchers in recent years. Kurmi et al. [87] applied the dropout regularization in adversarial training for feature alignment. The approach replaces the point estimates with distribution estimates, which increases the variance of the sample-based distribution and uses the corresponding inverse gradients to align the source and target domain features. The point estimates are obtained by a single discriminator, and the distribution estimates are obtained by a Monte Carlo dropout discriminator. Saito et al. [82] aligned the source and target domain distribution by reducing the decision boundary. In terms of feature space transformation, Hoffman et al. [86] proposed the cycle-consistent adversarial domain adaptation (CyCADA), which combines the ideas of adversarial training and feature space transformation. They trained the model on multiple loss functions while performing feature space and pixel space alignment. Finally, the TL is completed by combining the cyclic consistency loss with the adversarial loss.

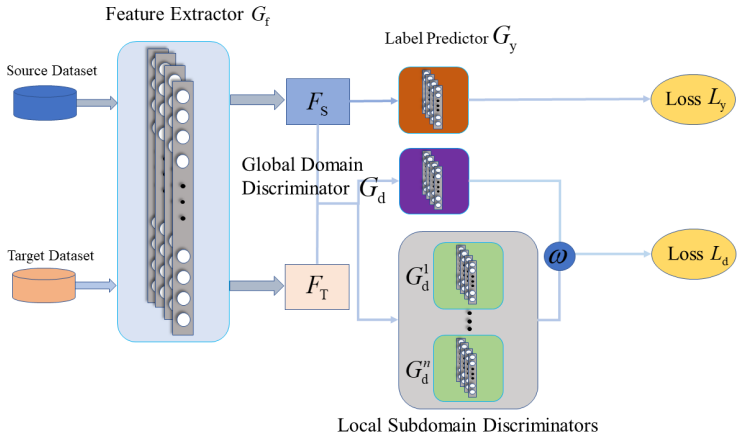

Figure 12. Architecture of DAAN [79]. Local domain discriminators $G_d^1(\cdot), \ldots, G_d^n(\cdot)$ are added to DANN. Note that ω is a dynamic measurement factor which measures the weights of $G_d(\cdot)$ and $G_d^1(\cdot), \ldots, G_d^n(\cdot)$.

The feature transformation approach of ATL that integrates the gradient inversion layer and domain classifier into a deep network has also achieved great results in the field of target detection. Regions with CNN features (R-CNN) [109] is the first model to apply deep learning to target detection successfully. Chen et al. [88] proposed the Faster R-CNN model based on R-CNN to solve the domain offset by training two domain adaptation components through adversarial learning. He et al. [89] was inspired by Faster R-CNN to propose the multi-adversarial Faster-RCNN (MAF) for accurate object detection. MAF uses a multi adversarial domain classifier to design a feature obfuscation layer by layer domain and proposes an information invariant scale reduction module (SRM) for hierarchical feature map resizing. The approach improves the training efficiency of adversarial domain adaptation. Recently, Xu et al. [90] proposed a classification regularization framework to solve the domain discrepancy problem in target detection by matching key image regions and important instances. The ideas of GRL and adversarial training were used to construct the regularization framework.

4.3. Other Approaches of ATL

In this subsection, other approaches of ATL are introduced. Dumoulin et al. [106] proposed adversarially learned inference (ALI), which simultaneously learns the bidirectional mapping between the feature space and data space; see Figure 13. The generator completes the feature space to data space mapping, the encoder learns the reverse mapping, and the discriminator discriminates the data from the bi-directional mapping to complete the adversarial training. Furthermore, using bi-directional generative networks is the bidirectional generative domain adaptive model proposed by Yang et al. [110] for the unsupervised TL, which completes cross-domain training by interpolating two intermediate domains to bridge the source and target domain. In addition to the simultaneous mapping of data and features, Ma et al. [107] proposed the graph convolutional adversarial network (GCAN) for unsupervised domain adaptation, which jointly models data structures, domain labels, and class labels in a deep framework. GCAN is designed with three effective alignment mechanisms, including structure-aware alignment, domain alignment, and center-of-mass alignment, to effectively learn domain invariance and semantic representation to reduce domain differences.

Kang et al. [108] explicitly modeled intra-class and inter-class domain differences for adversarial training, minimizing intra-class domain differences to avoid misalignment and maximizing inter-class domain differences to enhance the generalization ability of the model. Robbiano et al. [111] proposed the adversarial branching architecture search (ABAS) for unsupervised domain adaptation, which was the first time that a neural architecture search was introduced in unsupervised domain adaptation. Wang et al. [112] measured the confidence of the optimized model by the entropy of the model prediction, in which the adversarial training of domain adaptation was accomplished by minimizing the entropy. Mitsuzumi et al. [113] proposed a general representation of the unsupervised domain adaptation, generalized domain adaptation (GDA) [113], which can learn class invariant representations and domain adversarial classifiers without using any domain labels. In addition, Sun et al. [114] proposed a robust integrated network (REN) containing a teacher network and a student network for unsupervised TL.

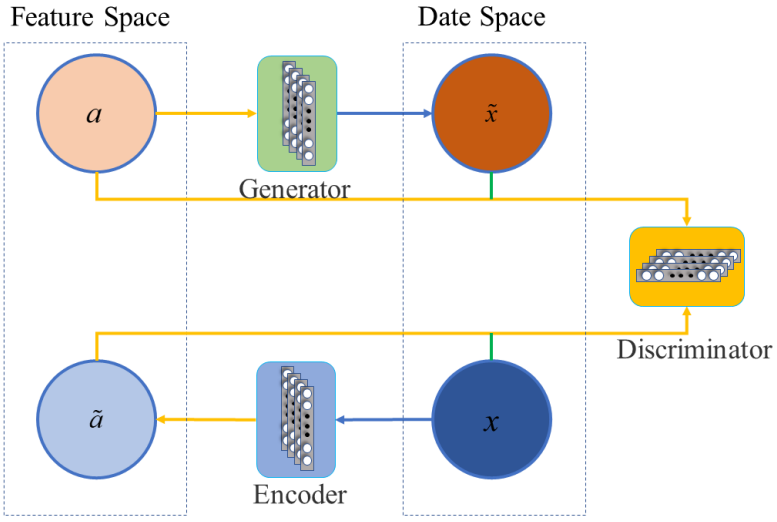

Figure 13. BiGAN simultaneously learns data space and feature space [106]. Here, $\boldsymbol{x}$ represents samples of the true data distribution in the data space, $\boldsymbol{a}$ represents samples of the true data distribution in the feature space, $\tilde{\boldsymbol{x}}$ represents samples of the generator (green) output, and $\tilde{\boldsymbol{a}}$ represents samples of the encoder (blue) output.

5. Relational-Based DTL

Relational-based DTL explores the relationships between samples in the source and target domains for cross-domain learning. It involves two mechanisms: reusing the source-domain relationship in a target domain (intra-domain relationship) and constructing a

cross-dataset relationship (inter-domain relationship). The Markov logic network (MLN) provides an ideal tool for reusing the source-domain relationship in a target domain, which is a representative approach of the former type. Davis et al. [93] proposed second-order MLN for DTL by extending the first-order MLN. The basic idea is to discover structural laws in the source domain by the Markov logic formulas with relational variables, and the relationship from the target domain is used to instantiate these formulas. After that, Haaren et al. [94] optimized the second-order formulas by directly computing the posterior distribution of second-order formulas, which is taken as the prior distribution of the second-order formulas in the target domain.

Another effective approach of relational-based DTL is constructing a cross-dataset samples relationship. This explores the relationship of different datasets in the source and target domains. Recently, Isobe et al. [91] proposed a collaborative learning framework for the single-source domain and multi-target domains. The teacher network is not a single system but a source domain with all target domains corresponding to n networks. Teacher networks learn different pixel-level classification capabilities by taking advantage of the differences existing in each domain, and the knowledge learned by the different teacher networks is integrated to obtain a network with more powerful generalization capabilities. The implications of the student networks are to make the teacher networks more closely connected by regularizing the weights of the teacher networks. Thereafter, He et al. [92] applied the collaborative learning approach on semantic segmentation tasks to exploit the essential semantic information across source domains.

In addition, open-set TL [115] considers the correspondence between the source and target domain categories, so open-set TL has also been applied in relational-based DTL.

6. Extensions and Additions

6.1. TL in Other Fields

With the rise of various machine learning approaches and the increasing demand for tasks, DTL is no longer satisfied with general classification and regression tasks. Approaches for combining with other machine learning have attracted attention and become frontier research in DTL. These approaches provide solutions for breaking data barriers, securing data, addressing sample shortages, increasing model arithmetic, and explaining deep learning models. This section gives a brief introduction to frontier advances of DTL. Table 5 summarizes various popular TL approaches in other fields.

Table 5. A brief summary of TL in other fields.

TL Approaches	Objective
Federated TL [116,117]	Protect the privacy of tasks data when multiple tasks are working together
Safe TL [118]	Reduce the aggressiveness inherited from the pre-trained model
Few-Shot TL [119,120]	Enhance the association of few labeled samples with unlabeled samples
Open Set TL [121]	Solve the problem of inconsistent source and target domain categories
Lifelong TL [122]	Use TL techniques to improve the effectiveness of lifelong learning Adaptively
Reinforcement TL [123,124]	Reduce the interference of environmental changes on reinforcement learning

Although federated learning (FL) is an effective means to break data barriers, FL requires each client to collect its local data independently and thus forms different source domains, resulting in the weak generalization ability of the model. Zhang et al. [117]

jointed the adversarial learning approach to solve the weak generalization, which measures and aligns the distributions between different source domains by matching each distribution to a reference distribution. For data security issues, Zhang et al. [116] proposed privacy-preserving TL for data security issues to avoid information leakage when generalizing the domain, which provides security for FL under data isolation. For the safe TL, Zhang et al. [118] applied a related model slicing technique. The approach dramatically improves the transfer accuracy by reducing the defect inheritance during TL while retaining the useful knowledge of the original model. Huang et al. [119] improved few-shot learning through TL. The spatial relationships of local descriptions, which were ignored in previous few-shot learning approaches, are effectively considered to make the learned image similarities better serve the desired domain alignment. In addition, few-shot learning is also combined with meta-learning for domain adaption. Cheng et al. [120] solved the problem of few-shot learning in which the base class and new class data come from different domains by combining meta-learning strategies. In the field of open set TL, Zhuang et al. [121] proposed a self-supervised discovery adapter to discover the implicit classes in both domains and determine the correspondence between the other categories using a part of the categories shared by both domains. In the field of lifelong TL, Yao et al. [122] proposed an adversarial feature alignment approach to address the catastrophic forgetting phenomenon, which focuses on incremental multitask image classification scenarios to provide a solution to the phenomenon in lifelong learning. In the field of reinforcement learning, Driessel et al. [123] introduced TL to transfer parameters of reinforcement learning. This work accomplishes parameter transfer for reinforcement learning by freezing the internal dynamics of learning and the value function. The TL, combined with other machine learning approaches, can utilize the characteristics of each machine learning approach to transfer knowledge in a targeted way and solve problems that cannot be solved in the original learning approach. The approach has a strong potential as a frontier and hot research area in machine learning.

6.2. Recent Advances in UTL

Less labeling or no labeling in the dataset has been a complex problem in machine learning and practical applications. Unsupervised transfer learning (UTL) has been an interesting area of TL. Many effective UTL approaches have been described in previous articles, such as self-training and ATL. This section adds to the recent progress of UTL approaches.

With the development of deep learning, the dimensionality reduction and clustering approaches in unsupervised learning have been optimized by deep learning. For completely unsupervised learning, where both the source and target domains lack labeling, Menapace et al. [125] proposed domain-independent deep clustering models by constructing data collected from multiple source domains. Some scholars have proposed the idea of joint learning by combining the approaches of clustering and dimensionality reduction with other approaches. Tian et al. [126] introduced local manifold learning for TL by combining clustering, center matching, and self-learning. The approach achieved good performance on both unsupervised and semi-supervised learning. For semi-supervised TL, Deng et al. [127], inspired by joint learning, proposed joint clustering and discriminative feature alignment (JCDFA). JCDFA unifies the mining of discriminative features and alignment of class discriminative features into a single framework to solve the discriminative clustering task for unlabeled target domains.

In addition, Luo et al. [128] explored the knowledge transfer mechanism and proposed a conditional kernel Bures (CKB) metric for characterizing differences in conditional distributions to learn the conditional invariant and discriminative features of UTL. Huang et al. [129] proposed the effective label propagation (ELP) to solve the semi-supervised TL, which enhances inter-domain semantic consistency through cyclic discrepancy loss and enhances intra-domain semantic information propagation through a self-training strategy to improve the feature discriminability in the target domain. Sun et al. [130] trained a prediction model by choosing hierarchical generation and decoupling approaches

within the framework of a variational auto-encoder, which can be generalized to new domains. In addition, for semi-supervised TL, Sharma et al. [131] proposed instance-level affinity-based domain adaption (ILA-DA) to extract similar and different samples across domains by using a multi-sample comparison loss to drive the domain alignment process. ILA-DA considers both intra-class clustering and inter-class separation to reduce the boundary noise of the classifier. In addition, Wu et al. [132] improved the UTL entropy minimization by introducing diversity maximization to regulate entropy. UTL is the most common problem in practical applications, which deserves further attention by researchers.

7. Summary and Future Prospects

In this paper, we have reviewed the development of deep transfer learning (DTL) in the past decades and summarized the related mechanisms and strategies. According to their models, functions, and operation objects, we have classified DTL into four categories and further divided them into subcategories. In particular, we have demonstrated the representative models and summarized their contributions and weaknesses. Last but not least, we have given extensions and additions to DTL, which include the frontier concerns of TL and the recent advance of unsupervised TL. It is indicated that DTL has enormous advantages over traditional machine learning, and it has great potential for many real-world applications.

Although DTL has achieved great success, some essential directions need to be further investigated.

- Interpretability of DTL is a great challenge to be explored. In the field of deep learning, there is a lack of interpretability of the learning process due to the existence of black boxes. The problem continues in the DTL area, and the development of DTL requires further investigation of the interpretability.
- How to reduce the effects of negative transfer while transferring knowledge from source domains to target domains is also an important issue. Therefore, improving TL algorithms and making theoretical innovations to avoid negative transfer should be considered.
- The single DTL approach has weak ability in practical applications. Joint learning [117,118,123] and multi-view learning [133,134] can provide a good way to solve this problem. It is interesting to integrate these approaches with DTL.
- The current work summarizes existing approaches, and we will compare them on datasets/tasks and hopefully give a ranking.

Author Contributions: Conceptualization, F.Y., X.X. and Y.L.; methodology, F.Y., X.X. and Y.L.; software, F.Y., X.X. and Y.L.; validation, F.Y., X.X. and Y.L.; formal analysis, Y.L.; investigation, X.X.; resources, X.X.; data curation, F.Y.; writing—original draft preparation, F.Y., X.X. and Y.L.; writing—review and editing, F.Y., X.X. and Y.L.; visualization, X.X.; supervision, Y.L.; project administration, X.X.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Shanghai Agriculture Applied Technology Development Program under Grant X2022-02-08-00-12-F01164 and China Postdoctoral Science Foundation under Grant 2021M702078.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The open-source codes and datasets for the four categories of DTL approaches mentioned in this paper are listed in Table A1.

Table A1. A summary of approaches with open-source codes.

Categories	Subcategories	Solvers	Dataset	Open-Source
Model-Based	Fine-Tuning	Fine-Tune Layer-by-Layer Yosinski et al. [45]	ImageNet	http://yosinski.com/transfer
		Improve Regularization and Robustness Li et al. [50]	ImageNet	https://github.com/NEU-StatsML-Research/Regularized-Self-Labeling
	Self-Training	Discover the Limitations of Pre-Training He et al. [51]	ImageNet, COCO	https://github.com/facebookresearch/detectron
		Self-train with Noisy Xie et al. [52]	ImageNet	https://github.com/google-research/noisystudent
	Transferring Attention Transformer	Image Processing Transformer (IPT) Chen et al. [54]	ImageNet	https://github.com/huawei-noah/Pretrained-IPT
		Bert Pre-training of Image Transformers (BEiT) Bao et al. [55]	ImageNet	https://github.com/microsoft/unilm/tree/master/beit
		Transferable Vision Transformer (TVT) Yang et al. [57]	Office-31	https://github.com/uta-smile/TVT
		Cross-domain Transformer (CDTrans) Xu et al. [58]	VisDA-2017, Office-Home, Office-31, DomainNet	https://github.com/cdtrans/cdtrans
Discrepancy-Based	Dual-Stream Architecture	Deep Correlation Alignment Sun et al. [68]	Office	https://github.com/VisionLearningGroup/CORAL
		Deep Domain Confusion (DDC) Tzeng et al. [60]	Office-31	https://github.com/erlendd/ddan
		Deep Adaptation Networks (DAN) Long et al. [62]	Office-31	http://github.com/thuml/DAN
		Joint adaptation networks (JAN) Long et al. [64]	Office-31, ImageCLEF-DA	http://github.com/thuml/JAN
		Deep Subdomain Adaptation Network (DSAN) Zhu et al. [63]	Office-31, ImageCLEF-DA, Office-Home, VisDA-2017	https://github.com/easezyc/deep-transfer-learning

Table A1. *Cont.*

Categories	Subcategories	Solvers	Dataset	Open-Source
		Dynamic Distribution Adaptation Network (DDAN) Wang et al. [65]	USPS+MNIST, Amazon review, Office-31, ImageCLEF-DA, Office-Home	http://transferlearning.xyz
		Central Moment Discrepancy (CMD) Zellinger et al. [66]	Amazon review, Office	https://github.com/wzell/cmd
	Operating on Image Features	Sample-to-Sample Self-Distillation Yoon et al. [74]	Office-Home, DomainNet	https://github.com/userb2020/s3d
		Learning to Match (L2M) Yu et al. [75]	ImageCLEF-DA, Office-Home, VisDA2017, Office-31	https://github.com/jindongwang/transferlearning
		Style-Agnostic Networks (SagNets) Nam et al. [73]	Office-Home, PACS, DomainNet	https://github.com/hyeonseobnam/sagnet
GAN-Based	Feature Extraction	Domain-Adversarial Training of Neural Networks (DANN) Ganin et al. [76]	Office	https://github.com/ddtm/caffe/tree/grl
		Conditional Adversarial Domain Adaptation (CADN) Long et al. [77]	Office-31, ImageCLEF-DA, Office-Home, Digits, VisDA-2017	http://github.com/thuml/CDAN
		Multi-adversarial Domain Adaptation (MADA) Pei et al. [78]	Office-31, ImageCLEF-DA	http://github.com/thuml/MADA
		Dynamic Adversarial Adaptation Network (DAAN) Yu et al. [79]	ImageCLEF-DA, Office-Home	http://transferlearning.xyz
		Cycle-Consistent Adversarial Networks (CycleGAN) Zhu et al. [81]	ImageNet	https://github.com/junyanz/CycleGAN
	Feature Transformation	Maximum Classifier Discrepancy (MCD) Saito et al. [82]	Digits, VisDA, Toy	https://github.com/mil-tokyo/MCD_DA
		Adversarial Feature Augmentation Volpi et al. [83]	SVHN, MNIST, NYUD	https://github.com/ricvolpi/adversarial-feature-augmentation
		Adversarial Discriminative Domain Adaptation (ADDA) Tzeng et al. [84]	SVHN, MNIST, USPS	https://github.com/thuml/Transfer-Learning-Library
		Simulated Generative Adversarial Networks (SimGAN) Shrivastava et al. [85]	MPIIGaze	https://github.com/0b01/SimGAN-Captcha
		Curriculum based Dropout Discriminator for Domain Adaptation (CD3A) Kurmi et al. [87]	ImageCLE, Office-31, Office-Home	https://github.com/DelTA-Lab-IITK/CD3A

Table A1. *Cont.*

Categories	Subcategories	Solvers	Dataset	Open-Source
		Region Proposal Network (RPN) Chen et al. [88]	Cityscapes ,KITTI, SIM10K	https://github.com/yuhuayc/da-faster-rcnn
		Categorical Consistency Regularization (CCR) Xu et al. [90]	Cityscapes, Foggy Cityscapes, BDD100k, PASCAL, VOC, Clipart1k	https://github.com/Megvii-Nanjing/CR-DA-DET
Relational-Based	Cross-Domain Relationship	Collaborative Consistency Learning (CCL) Isobe et al. [91]	GTA5, SYNTHIA, Cityscapes, Mapillary	https://github.com/junpan19/MTDA

References

1. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 815–823.
2. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. Deepface: Closing the Gap to Human-Level Performance in Face Verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 1701–1708.
3. Geiger, A.; Lenz, P.; Urtasun, R. Are We Ready for Autonomous Driving? The Kitti Vision Benchmark Duite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
4. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. CheXNET: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv* **2017**, arXiv:1711.05225
5. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
6. Wang, Z. Theoretical Guarantees of Transfer Learning. *arXiv* **2018**, arXiv:1810.05986
7. Vural, E. Generalization Bounds for Domain Adaptation Via Domain Transformations. In Proceedings of the 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP), Providence, RI, USA, 16–21 June 2021 **2018**.
8. Wang, W.; Li, B.; Yang, S.; Sun, J.; Ding, Z.; Chen, J.; Dong, X.; Wang, Z.; Li, H. A Unified Joint Maximum Mean Discrepancy for Domain Adaptation. *arXiv* **2021**, arXiv:2101.09979.
9. Galanti, T.; Benaim, S.; Wolf, L. Risk Bounds for Unsupervised Cross-Domain Mapping with IPMs. *J. Mach. Learn. Res.* **2021**, *22*, 90–91.
10. Phung, T.; Le, T.; Vuong, T.L.; Tran, T.; Tran, A.; Bui, H.; Phung, D. On Learning Domain-Invariant Representations for Transfer Learning with Multiple Sources. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 27720–27733.
11. Teshima, T.; Sato, I.; Sugiyama, M. Few-Shot Domain Adaptation by Causal Mechanism Transfer. In Proceedings of the PMLR: International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 9458–9469.
12. Wu, X.; Manton, J.H.; Aickelin, U.; Zhu, J. An Information-Theoretic Analysis for Transfer Learning: Error Bounds and Applications. *arXiv* **2022**, arXiv:2207.05377.
13. Acuna, D.; Zhang, G.; Law, M.T.; Fidler, S. f-Domain-Adversarial Learning: Theory and Algorithms. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual, 18–24 July 2021; pp. 66–75.
14. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
15. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2020**, *109*, 43–76. [CrossRef]
16. Wang, J.; Chen, Y. *Introduction to Transfer Learning*; Electronic Industry Press: Beijing, China, 2021.
17. Yang, Q; Zhang, Y. *Transfer Learning*; Artificial Intelligence Technology Series, Machinery Industry Press: Beijing, China, 2020.
18. Zadrozny, B. Learning and Dvaluating Classifiers under Sample Selection Bias. In Proceedings of the Twenty-First International Conference on Machine Learning, Banff Alberta, AB, Canada, 4–8 July 2004; p. 114.
19. Dai, W.; Yang, Q.; Xue, G.R.; Yu, Y. Boosting for Transfer Learning. In Proceedings of the ICML '07, 24th International Conference on Machine Learning, Corvalis Oregon, OR, USA, 20–24 June 2007; Volume 227, pp. 193–200. [CrossRef]
20. Yao, Y.; Doretto, G. Boosting for Transfer Learning with Multiple Sources. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1855–1862.
21. Huang, J.; Gretton, A.; Borgwardt, K.; Schölkopf, B.; Smola, A. Correcting Sample Selection Bias by Unlabeled Data. *Adv. Neural Inf. Process. Syst.* **2006**, *19*, 601–608.
22. Pan, S.J.; Tsang, I.W.; Kwok, J.T.; Yang, Q. Domain Adaptation Via Transfer Component Analysis. *IEEE Trans. Neural Netw.* **2010**, *22*, 199–210. [CrossRef] [PubMed]

23. Long, M.; Wang, J.; Ding, G.; Sun, J.; Yu, P.S. Transfer Feature Learning with Joint Distribution Adaptation. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2200–2207.
24. Wang, J.; Feng, W.; Chen, Y.; Yu, H.; Huang, M.; Yu, P.S. Visual Domain Adaptation with Manifold Embedded Distribution Alignment. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Korea, 22–26 October 2018; pp. 402–410.
25. Gong, B.; Shi, Y.; Sha, F.; Grauman, K. Geodesic Flow Kernel for Unsupervised Domain Adaptation. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2066–2073.
26. Sun, B.; Feng, J.; Saenko, K. Return of Frustratingly Easy Domain Adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.
27. Fernando, B.; Habrard, A.; Sebban, M.; Tuytelaars, T. Unsupervised Visual Domain Adaptation Using Subspace Alignment. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2960–2967.
28. Fei-Fei, L.; Fergus, R.; Perona, P. One-Shot Learning of Object Categories. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 594–611. [CrossRef] [PubMed]
29. Bonilla, E.V.; Chai, K.; Williams, C. Multi-Task Gaussian Process Prediction. *Adv. Neural Inf. Process. Syst.* **2007**, *20*, 601–608.
30. Yang, J.; Yan, R.; Hauptmann, A.G. Cross-Domain Video Concept Detection Using Adaptive SVMs. In Proceedings of the 15th ACM International Conference on Multimedia, Augsburg, Germany, 25–29 September 2007; pp. 188–197.
31. Aytar, Y.; Zisserman, A. Tabula Rasa: Model Transfer for Object Category Detection. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2252–2259.
32. Mihalkova, L.; Huynh, T.; Mooney, R.J. Mapping and Revising Markov Logic Networks for Transfer Learning. In Proceedings of the AAAI, Vancouver, BC, Canad, 22–26 July 2007; Volume 7, pp. 608–614.
33. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 84–90. [CrossRef]
34. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556
35. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2016; pp. 1–9.
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Ecognition, Boston, MA, USA, 8–10 June 2016; pp. 770–778.
37. Csurka, G.; Baradel, F.; Chidlovskii, B.; Clinchant, S. Discrepancy-Based Networks for Unsupervised Domain Adaptation: A Comparative Study. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 2630–2636.
38. Li, D.; Yang, Y.; Song, Y.Z.; Hospedales, T.M. Deeper, Broader and Artier Domain Generalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5542–5550.
39. Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; Darrell, T. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In Proceedings of the International Conference on Machine Learning. PMLR, Bejing, China, 22–24 June 2014; pp. 647–655.
40. Saxena, S.; Verbeek, J. Heterogeneous Face Recognition with CNNs. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–10 October 2016; Springer: New York, NY, USA, 2016; pp. 483–491.
41. Fernando, B.; Tommasi, T.; Tuytelaars, T. Joint Cross-Domain Classification and Subspace Learning for Unsupervised Adaptation. *Pattern Recognit. Lett.* **2015**, *65*, 60–66. [CrossRef]
42. Long, M.; Wang, J.; Ding, G.; Sun, J.; Yu, P.S. Transfer Joint Matching for Unsupervised Domain Adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 1410–1417.
43. Chopra, S.; Balakrishnan, S.; Gopalan, R. DLID: Deep Learning for Domain Adaptation by Interpolating between Domains. In Proceedings of the ICML Workshop on Challenges in Representation Learning, Atlanta, GA, USA, 16–21 June 2013; Volume 2.
44. Wang, M.; Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing* **2018**, *312*, 135–153. [CrossRef]
45. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How Transferable are features in deep neural networks? *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 3320–3328.
46. Chu, B.; Madhavan, V.; Beijbom, O.; Hoffman, J.; Darrell, T. Best Practices for Fine-tuning Visual Classifiers to New Domains. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–10 October 2016; Springer: New York, NY, USA, 2016; pp. 435–442.
47. Rozantsev, A.; Salzmann, M.; Fua, P. Beyond Sharing Weights for Deep Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 801–814. [CrossRef]
48. Rozantsev, A.; Salzmann, M.; Fua, P. Residual Parameter Transfer for Deep Domain Adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018; pp. 4339–4348.
49. Guo, Y.; Shi, H.; Kumar, A.; Grauman, K.; Rosing, T.; Feris, R. Spottune: Transfer Learning through Adaptive Fine-tuning. In Proceedings of the IEEE/CVF Conference on Cmputer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4805–4814.
50. Li, D.; Zhang, H. Improved Regularization and Robustness for Fine-tuning in Neural Networks. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, [CrossRef]

51. He, K.; Girshick, R.; Dollár, P. Rethinking Imagenet Pre-Training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27October–2 November 2019; pp. 4918–4927.
52. Xie, Q.; Luong, M.T.; Hovy, E.; Le, Q.V. Self-Training with Noisy Student Improves Imagenet Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 10687–10698.
53. Zoph, B.; Ghiasi, G.; Lin, T.Y.; Cui, Y.; Liu, H.; Cubuk, E.D.; Le, Q. Rethinking Pre-Training and Self-Training. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3833–3845.
54. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-Trained Image Processing Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 12299–12310.
55. Bao, H.; Dong, L.; Wei, F. Beit: Bert Pre-Training of Image Transformers. *arXiv* **2021**, arXiv:2106.08254.
56. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
57. Yang, J.; Liu, J.; Xu, N.; Huang, J. TVT: Transferable Vision Transformer for Unsupervised Domain Adaptation. *arXiv* **2021**, arXiv:2108.05988.
58. Xu, T.; Chen, W.; Wang, P.; Wang, F.; Li, H.; Jin, R. Cdtrans: Cross-Domain Transformer for Unsupervised Domain Adaptation. *arXiv* **2021**, arXiv:2109.06165.
59. Ghifary, M.; Kleijn, W.B.; Zhang, M. Domain Adaptive Neural Networks for Object Recognition. In Proceedings of the Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, 1–5 December 2014; Springer: New York, NY, USA, 2014; pp. 898–904.
60. Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; Darrell, T. Deep Domain Confusion: Maximizing for Domain Invariance. *arXiv* **2014**, arXiv:1412.3474.
61. Tzeng, E.; Hoffman, J.; Darrell, T.; Saenko, K. Simultaneous Deep Transfer Across Domains and Tasks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 4068–4076.
62. Long, M.; Cao, Y.; Wang, J.; Jordan, M. Learning Transferable Features with Deep Adaptation Networks. In Proceedings of the PMLR: International Conference on Machine Learning, Lille, France, 1–9 July 2015; pp. 97–105.
63. Zhu, Y.; Zhuang, F.; Wang, J.; Ke, G.; Chen, J.; Bian, J.; Xiong, H.; He, Q. Deep Subdomain Adaptation Network for Image Classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 1713–1722. [CrossRef]
64. Long, M.; Zhu, H.; Wang, J.; Jordan, M.I. Deep Transfer Learning with Joint Adaptation Networks. In Proceedings of the International Conference on Machine Learning. PMLR, Sydney, Australia, 6–11 August 2017; pp. 2208–2217.
65. Wang, J.; Chen, Y.; Feng, W.; Yu, H.; Huang, M.; Yang, Q. Transfer Learning with Dynamic Distribution Adaptation. *ACM Trans. Intell. Syst. Technol. (TIST)* **2020**, *11*, 1–25. [CrossRef]
66. Zellinger, W.; Grubinger, T.; Lughofer, E.; Natschläger, T.; Saminger-Platz, S. Central Moment Discrepancy (CMD) for Domain-Invariant Representation Learning. *arXiv* **2017**, arXiv:1702.08811.
67. Zhu, Y.; Zhuang, F.; Wang, J.; Chen, J.; Shi, Z.; Wu, W.; He, Q. Multi-Representation Adaptation Network for Cross-Domain Image Classification. *Neural Netw.* **2019**, *119*, 214–221. [CrossRef]
68. Sun, B.; Saenko, K. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–10 October 2016; Springer: New York, NY, USA, 2016; pp. 443–450.
69. Li, Y.; Wang, N.; Shi, J.; Hou, X.; Liu, J. Adaptive Batch Normalization for Practical Domain Adaptation. *Pattern Recognit.* **2018**, *80*, 109–117. [CrossRef]
70. Shen, J.; Qu, Y.; Zhang, W.; Yu, Y. Wasserstein Distance Guided Representation Learning for Domain Adaptation. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
71. Xu, R.; Liu, P.; Wang, L.; Chen, C.; Wang, J. Reliable Weighted Optimal Transport for Unsupervised Domain Adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 4394–4403.
72. Damodaran, B.B.; Kellenberger, B.; Flamary, R.; Tuia, D.; Courty, N. Deepjdot: Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 447–463.
73. Nam, H.; Lee, H.; Park, J.; Yoon, W.; Yoo, D. Reducing Domain Gap by Reducing Style Bias. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 8690–8699.
74. Yoon, J.; Kang, D.; Cho, M. Semi-Supervised Domain Adaptation via Sample-to-Sample Self-Distillation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 1978–1987.
75. Yu, C.; Wang, J.; Liu, C.; Qin, T.; Xu, R.; Feng, W.; Chen, Y.; Liu, T.Y. Learning to Match Distributions for Domain Adaptation. *arXiv* **2020**, arXiv:2007.10791.
76. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
77. Long, M.; Cao, Z.; Wang, J.; Jordan, M.I. Conditional Adversarial Domain Adaptation. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, [CrossRef]
78. Pei, Z.; Cao, Z.; Long, M.; Wang, J. Multi-Adversarial Domain Adaptation. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.

79. Yu, C.; Wang, J.; Chen, Y.; Huang, M. Transfer Learning with Dynamic Adversarial Adaptation Network. In Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM), Beijing, China, 08–11 November 2019; pp. 778–786.
80. Xu, M.; Zhang, J.; Ni, B.; Li, T.; Wang, C.; Tian, Q.; Zhang, W. Adversarial Domain Ddaptation with Domain Mixup. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 6502–6509.
81. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
82. Saito, K.; Watanabe, K.; Ushiku, Y.; Harada, T. Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3723–3732.
83. Volpi, R.; Morerio, P.; Savarese, S.; Murino, V. Adversarial Feature Augmentation for Unsupervised Domain Adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5495–5504.
84. Tzeng, E.; Hoffman, J.; Saenko, K.; Darrell, T. Adversarial Discriminative Domain Adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 7167–7176.
85. Shrivastava, A.; Pfister, T.; Tuzel, O.; Susskind, J.; Wang, W.; Webb, R. Learning from Simulated and Unsupervised Images through Adversarial Training. In Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 2107–2116.
86. Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.Y.; Isola, P.; Saenko, K.; Efros, A.; Darrell, T. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In Proceedings of the PMLR: International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1989–1998.
87. Kurmi, V.K.; Subramanian, V.K.; Namboodiri, V.P. Exploring Dropout Discriminator for Domain Adaptation. *Neurocomputing* **2021**, *457*, 168–181. [CrossRef]
88. Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; Van Gool, L. Domain Adaptive Faster R-CNN for Object Detection in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3339–3348.
89. He, Z.; Zhang, L. Multi-Adversarial Faster-RCNN for Unrestricted Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6668–6677.
90. Xu, C.D.; Zhao, X.R.; Jin, X.; Wei, X.S. Exploring Categorical Regularization for Domain Adaptive Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 11724–11733.
91. Isobe, T.; Jia, X.; Chen, S.; He, J.; Shi, Y.; Liu, J.; Lu, H.; Wang, S. Multi-Target Domain Adaptation with Collaborative Consistency Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 8187–8196.
92. He, J.; Jia, X.; Chen, S.; Liu, J. Multi-Source Domain Adaptation with Collaborative Learning for Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 11008–11017.
93. Davis, J.; Domingos, P. Deep Transfer via Second-Order Markov Logic. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 217–224.
94. Van Haaren, J.; Kolobov, A.; Davis, J. TODTLER: Two-Order-Deep Transfer Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 29.
95. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078.
96. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. *Adv. Neural Inf. Process. Syst.* **2014**, *27* , 3104–3112.
97. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2014**, arXiv:1409.0473.
98. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Korea, 27 October–2 November 2019; pp. 3146–3154.
99. Das, T.; Bruintjes, R.J.; Lengyel, A.; van Gemert, J.; Beery, S. Domain Adaptation for Rare Classes Augmented with Synthetic Samples. *arXiv* **2021**, arXiv:2110.12216.
100. Xu, Z.; Lee, G.H.; Wang, Y.; Wang, H. Graph-Relational Domain Adaptation. *arXiv* **2022**, arXiv:2202.03628.
101. Li, X.; Dai, Y.; Ge, Y.; Liu, J.; Shan, Y.; Duan, L.Y. Uncertainty Modeling for Out-of-Distribution Generalization. *arXiv* **2022**, arXiv:2202.03958.
102. Ghifary, M.; Kleijn, W.B.; Zhang, M.; Balduzzi, D. Domain Generalization for Object Recognition with Multi-Task Autoencoders. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 2551–2559.
103. Zhu, Y.; Zhuang, F.; Wang, D. Aligning Domain-Specific Distribution and Classifier for Cross-domain Classification from Multiple Dources. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 5989–5996.
104. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 539–546.

105. Borgwardt, K.M.; Gretton, A.; Rasch, M.J.; Kriegel, H.P.; Schölkopf, B.; Smola, A.J. Integrating Structured Biological Data by Kernel Maximum Mean Discrepancy. *Bioinformatics* **2006**, *22*, e49–e57. [CrossRef] [PubMed]
106. Dumoulin, V.; Belghazi, I.; Poole, B.; Mastropietro, O.; Lamb, A.; Arjovsky, M.; Courville, A. Adversarially Learned Inference. *arXiv* **2016**, arXiv:1606.00704.
107. Ma, X.; Zhang, T.; Xu, C. GCAN: Graph Convolutional Adversarial Network for Unsupervised Domain Adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Korea, 27 October–2 November 2019; pp. 8266–8276.
108. Kang, G.; Jiang, L.; Yang, Y.; Hauptmann, A.G. Contrastive Adaptation Network for Unsupervised Domain Adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Korea, 27 October–2 November 2019; pp. 4893–4902.
109. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
110. Yang, G.; Xia, H.; Ding, M.; Ding, Z. Bi-Directional Generation for Unsupervised Domain Adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 6615–6622.
111. Robbiano, L.; Rahman, M.R.U.; Galasso, F.; Caputo, B.; Carlucci, F.M. Adversarial Branch Architecture Search for Unsupervised Domain Adaptation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, New Orleans, LA, USA, 21–24 June 2022; pp. 2918–2928.
112. Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; Darrell, T. Tent: Fully Test-Time Adaptation by Entropy Minimization. *arXiv* **2020**, arXiv:2006.10726.
113. Mitsuzumi, Y.; Irie, G.; Ikami, D.; Shibata, T. Generalized Domain Adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 1084–1093.
114. Sun, H.; Lin, L.; Liu, N.; Zhou, H. Robust Ensembling Network for Unsupervised Domain Adaptation. In Proceedings of the Pacific Rim International Conference on Artificial Intelligence, Hanoi, Vietnam, 8–12 November 2021; Springer: New York, NY, USA, 2021; pp. 530–543.
115. Panareda Busto, P.; Gall, J. Open Set Domain Adaptation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 754–763.
116. Zhang, B.; Chen, C.; Wang, L. Privacy-preserving Transfer Learning via Secure Maximum Mean Discrepancy. *arXiv* **2020**, arXiv:2009.11680.
117. Zhang, L.; Lei, X.; Shi, Y.; Huang, H.; Chen, C. Federated Learning with Domain Generalization. *arXiv* **2021**, arXiv:2111.10487.
118. Zhang, Z.; Li, Y.; Wang, J.; Liu, B.; Li, D.; Guo, Y.; Chen, X.; Liu, Y. ReMoS: Reducing Defect Inheritance in Transfer Learning via Relevant Model Slicing. In Proceedings of the 2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE), Pittsburgh, PA, USA, 21–29 May 2022; pp. 1856–1868.
119. Huang, S.; Yang, W.; Wang, L.; Zhou, L.; Yang, M. Few-shot Unsupervised Domain Adaptation with Image-to-class Sparse Similarity Encoding. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 677–685.
120. Cheng, Y.C.; Lin, C.S.; Yang, F.E.; Wang, Y.C.F. Few-Shot Classification in Unseen Domains by Episodic Meta-Learning Across Visual Domains. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 434–438.
121. Zhuang, J.; Chen, Z.; Wei, P.; Li, G.; Lin, L. Open Set Domain Adaptation by Novel Class Discovery. *arXiv* **2022**, arXiv:2203.03329.
122. Yao, X.; Huang, T.; Wu, C.; Zhang, R.X.; Sun, L. Adversarial Feature Alignment: Avoid Catastrophic Forgetting in Incremental Task Lifelong Learning. *Neural Comput.* **2019**, *31*, 2266–2291. [CrossRef]
123. Van Driessel, G.; Francois-Lavet, V. Component Transfer Learning for Deep RL Based on Abstract Representations. *arXiv* **2021**, arXiv:2111.11525
124. Castagna, A.; Dusparic, I. Multi-Agent Transfer Learning in Reinforcement Learning-Based Ride-Sharing Systems. *arXiv* **2021**, arXiv:2112.00424
125. Menapace, W.; Lathuilière, S.; Ricci, E. Learning to Cluster Under Domain Shift. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: New York, NY, USA, 2020; pp. 736–752.
126. Tian, L.; Tang, Y.; Hu, L.; Ren, Z.; Zhang, W. Domain Adaptation by Class Centroid Matching and Local Manifold Self-Learning. *IEEE Trans. Image Process.* **2020**, *29*, 9703–9718. [CrossRef] [PubMed]
127. Deng, W.; Liao, Q.; Zhao, L.; Guo, D.; Kuang, G.; Hu, D.; Liu, L. Joint Clustering and Discriminative Feature Alignment for Unsupervised Domain Adaptation. *IEEE Trans. Image Process.* **2021**, *30*, 7842–7855. [CrossRef] [PubMed]
128. Luo, Y.W.; Ren, C.X. Conditional Bures Metric for Domain Adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 13989–13998.
129. Huang, Z.; Sheng, K.; Dong, W.; Mei, X.; Ma, C.; Huang, F.; Zhou, D.; Xu, C. Effective Label Propagation for Discriminative Semi-Supervised Domain Adaptation. *arXiv* **2020**, arXiv:2012.02621.
130. Sun, X.; Buettner, F. Hierarchical Domain Invariant Variational Auto-Encoding with Weak Domain Supervision. *arXiv* **2021**, arXiv:2101.09436.

131. Sharma, A.; Kalluri, T.; Chandraker, M. Instance Level Affinity-Based Transfer for Unsupervised Domain Adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 5361–5371.
132. Wu, X.; Zhang, S.; Zhou, Q.; Yang, Z.; Zhao, C.; Latecki, L.J. Entropy Minimization Versus Diversity Maximization for Domain Adaptation. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–12. [CrossRef] [PubMed]
133. Zhang, J.; Qi, L.; Shi, Y.; Gao, Y. More is Better: A Novel Multi-view Framework for Domain Generalization. *arXiv* **2021**, arXiv:2112.12329
134. Deng, Z.; Zhou, K.; Yang, Y.; Xiang, T. Domain Attention Consistency for Multi-Source Domain Adaptation. *arXiv* **2021**, arXiv:2111.03911